Open Research Online

The Open University's repository of research publications and other research outputs

Multimedia resource discovery

Book Section

How to cite:

Rüger, Stefan (2011). Multimedia resource discovery. In: Melucci, Massimo and Baeza-Yates, Ricardo eds. Advanced Topics in Information Retrieval. The Information Retrieval Series. Heidelberg: Springer, pp. 157–186.

For guidance on citations see FAQs.

© 2011 Stefan Rueger; Springer-Verlag

Version: Accepted Manuscript

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1007/978-3-642-20946-8_7

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Multimedia resource discovery[1]

Stefan Rüger
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes MK7 6AA, UK
s.rueger@open.ac.uk

August 13, 2010

[1]This book chapter is an updated re-print of Rüger (2009), Multimedia resource discovery, in Göker and Davies (eds), Information Retrieval: Searching in the 21st Century, pp 39–62, Wiley, with excerpts from Rüger (2010), Multimedia information retrieval, Lecture notes in the series Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan and Claypool Publishers, DOI: 10.2200/S00244ED1V01Y200912ICR010

0.1 Paradigms and challenges

Resource discovery is more than just search: it is browsing, searching, selecting, assessing and evaluating, ie, ultimately accessing information. Giving users access to collections is one of the defining tasks of a library. For thousands of years the traditional methods of resource discovery have been facilitated by librarians: they create reference cards with meta-data that are put into catalogues (nowadays, databases); they also place the objects in physical locations that follow certain classification schemes and they answer questions at the reference desk.

The advent of digital documents has radically changed the organisation principles; now it is possible to *automatically* index and search document collections as big as the world-wide web *à la* Google and browse collections utilising author-inserted links. It is almost as if automated processing has turned the traditional library access paradigm upside down. Instead of searching meta-data catalogues in order to retrieve the document, web search engines search the full content of documents and retrieve their meta-data, ie, the location where documents can be found. Not all manual intervention has been abandoned, though. For example, the Yahoo directory[1] is an edited classification scheme of submitted web sites that are put into a browsable directory structure akin to library classification schemes.

At its very core, multimedia information retrieval means the process of searching for and finding multimedia documents; the corresponding research field is concerned with building multimedia search engines. The intriguing bit about multimedia retrieval is that the query itself can be a multimedia excerpt: for example, if you walk around in pleasant Milton Keynes, you may stumble across the interesting building that is depicted in Figure 1.

Figure 1: Milton Keynes's Peace Pagoda

Would it not be nice if you could just take a picture with your mobile phone and send it to a service that matches your picture to their database and tells you more about the building? The service could reply with *"Built by the monks and nuns of the Nipponzan Myohoji, this was the first Peace Pagoda to be built in the western hemisphere and enshrines sacred relics of Lord Buddha. The Inauguration ceremony, on 21st September 1980, was presided over by the late most Venerable Nichidattsu Fujii, founder …"*[2]

[1] http://dir.yahoo.com/
[2] http://www.mkweb.co.uk/places_to_visit/displayarticle.asp?id=411 accessed Aug 2010

Given the much wider remit of multimedia search over just text search, and assuming we could perfectly search with queries that are “multimedia” itself, what could we do with multimedia search?

The previous example is an obvious application for tourism. There are also applications for advertising that so much seems to underpin the whole search industry: Snaptell Inc, is a startup company that specialises in mobile image search; their idea is that customers take pictures from print-media adverts, send them in and receive promotion or product information, vouchers and so on. For example, customers sending in a picture of the print poster that advertises a new movie receive an exclusive trailer, see showtimes of cinemas in the area and, in theory, could straight away phone to order tickets. One added benefit for advertisers is that they receive feedback as to where print adverts were noticed. Snaptell has since then specialised on recognising book, CD and DVD covers, from photographs that can then be bought with a few clicks.

Another considerable and obvious application is for medical image databases. When someone who suffers from shortness of breath consults doctors, they might wonder where they have seen the light shadow on the x-ray before. If computers were able to match significant, medically relevant patterns with those in the database, they could return data on these diagnosed cases, so the specialists can undertake an informed differential diagnosis using medical-image retrieval (Figure 2).

Figure 2: Medical-image retrieval (mock-up)

The common factor of the previous examples was that documents and queries can consist of various different media. Figure 3 takes this observation radically forward by looking at the full matrix of combining different query modes (columns) with document repository types (rows). Entry A in this matrix corresponds to a traditional text search engine; this deploys a completely different technology than Entry B, a system that allows you to express musical queries by humming a tune and that then plays the corresponding song. The three C entries in Figure 3 correspond to a multi-modal video search engine allowing search by emotion with example images and text queries, eg, *find me video shots of “sad” scenes using an image of a dilapidated castle and the text “vampire”*. In contrast to this, Entry D could be a search engine with a query text box that returns BBC Radio 4 discussions.

It is relatively easy to come up with a usage scenario for each of the matrix elements in Figure 3: for example, the image input speech output matrix element might be “given an X-ray image of a patient’s chest, retrieve dictaphone documents with a relevant spoken description of a matching diagnosis”. However, creating satisfying retrieval solutions is highly non-trivial and the main subject of the multimedia information retrieval discipline. This chapter summarises different basic technologies involved in these multimedia search modes. Not all combinations are equally useful, desirable or well researched, though: Entry E might be a query where you roar like a lion and hope to retrieve a wildlife documentary.

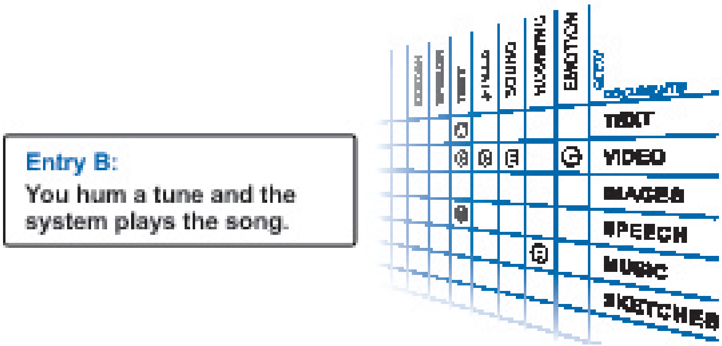


Figure 3: New search engine types

Undoubtedly, it is the automated approaches that have made all the difference to the way the vast world-wide web can be used. While the automated indexing of text documents has been successfully applied to collections as large as the world-wide-web for more than a decade now, multimedia indexing by content involves different, still less mature and less scalable technologies.

Multimedia collections pose their very own challenges; for example, images and videos don't often come with dedicated reference cards or meta-data, and when they do, as in museum collections, their creation will have been expensive and time-consuming. Section 0.3 explores the difficulties and limitations of automatically indexing, labelling and annotating image and video content. It briefly discusses the inherent challenges of the semantic gap, polysemy, fusion and responsiveness.

Even if all these challenges were solved, indexing sheer mass is no guarantee of a successful annotation either: While most of today's inter-library loan systems allow access to virtually any publication in the world — around 198m bibliographic records in OCLC's Worldcat database[3] in contrast to only 3m entries from Bowker's Books In Print that can be bought — students and researchers alike seem to be reluctant to actually make use of this facility. On the other hand, the much smaller catalogue offered by the online bookseller Amazon appears to be very popular, presumably owing to added services such as subject categories; fault tolerant search tools; personalised services telling the customer what's new in a subject area or what other people with a similar profile bought; pictures of book covers; media and customer reviews; access to the table of contents, to selections of the text and to the full-text index of popular books; and the perception of fast delivery. In the multimedia context Section 0.4 argues that automated added services such as visual queries, relevance feedback and summaries can prove useful for resource discovery in multimedia digital libraries. Sections 0.4.1 is about summarising techniques for videos, Section 0.4.2 exemplifies visualisation of search results, while Section 0.4.3 discusses content-based visual search modes such as query-by-example and relevance feedback.

Finally, Section 0.5 promotes browsing as resource discovery mode and looks at underlying techniques to automatically structure the document collection to support browsing.

[3]http://www.oclc.org/worldcat/statistics accessed Aug 2010

0.2 Basic Multimedia Search Technologies

The current best practice to index multimedia collections is via the generation of a library card, ie, a dedicated database entry of meta-data such as author, title, publication year and keywords. Depending on the concrete implementation these can be found with SQL queries, text-search engines or XML query language, but all these search modes are based on text descriptions of some form and are agnostic to the structure of the actual objects they refer to, be it books, CDs, videos, newspaper articles, paintings, sculptures, web pages, consumer products etc.

The text column of the matrix of Figure 3 is underpinned by text search technology and requires the textual representation of the multimedia objects, an approach that I like to call *piggy-back text retrieval*. Other approaches are based on an automatic classification of multimedia objects and on assigning words from a fixed vocabulary. This can be a certain camera motion that can be detected in a video (zoom, pan, tilt, roll, dolly in and out, truck left and right, pedestal up and down, crane boom, swing boom etc); a genre for music pieces such as jazz, classics; a generic scene description in images such as inside/outside, people, vegetation, landscape, grass, city-view etc or specific object detection like faces and cars etc. These approaches are known as *feature classification* or *automated annotation*.

The type of search that is most commonly associated with multimedia is *content-based*: The basic idea is that still images, music extracts, video clips themselves can be used as queries and that the retrieval system is expected to return 'similar' database entries. This technology differs most radically from the thousands-year-old library card paradigm in that there is no necessity for meta-data at all. In certain searches there is the desire to match not only the general type of scene or music that the the query represents, but instead one and only one exact multimedia object. For example, you take a picture of a painting in a gallery and submit this as a query to the gallery's catalogue in the hope of receiving the whole database record about this particular painting, and not a variant or otherwise similar exhibit. This is sometimes called *fingerprinting* or *known-item search*.

The rest of this section outlines these four basic multimedia search technologies.

0.2.1 Piggy-back text retrieval

Amongst all media types, TV video streams arguably have the biggest scope for automatically extracting text strings in a number of ways: directly from closed-captions, teletext or subtitles; automated speech recognition on the audio and optical character recognition for text embedded in the frames of a video. Full text search of these strings is the way in which most video retrieval systems operate, including Google's TV search engine http://video.google.com or Blinkx-TV http://www.blinkx.tv. In contrast to television, for which legislation normally requires subtitles to assist the hearing impaired, videos stored on DVD don't usually have textual subtitles. They have *subpicture* channels for different languages instead, which are overlayed on the video stream. This requires the extra step of optical character recognition, which can be done with a relatively low error rate owing to good quality fonts and clear background/foreground separation in the subpictures. In general, teletext has a much lower word error rate than automated speech recognition. In practice, it turns out that this does not matter too much as query words often occur repeatedly in the audio — the retrieval performance degrades gracefully with increased word error rates.

Web pages afford some context information that can be used for indexing multimedia objects. For example, words in the anchor text of a link to an image, a video clip or a music track, the file name of the object itself, meta-data stored within the files and other context information such as captions. A subset of these sources for text snippets are normally used in web image search engines.

Some symbolic music representations allow the conversion of music into text, such as MIDI files which contain a music representation in terms of pitch, onset times and duration of notes. By representing differences of successive pitches as characters one can, for example, map monophonic music to one-dimensional strings. A large range of different text matching techniques can be deployed, for example the edit distance of database strings with a string representation of a query. The edit distance between two strings computes the smallest number of deletions, insertions or character replacements that is necessary to transform one string into the other. In the case of query-by-humming, where a pitch tracker can convert the hummed query into a MIDI-sequence (Birmingham et al, 2006), the edit distance is also able to deal gracefully with humming errors. Other techniques create fixed-length strings, so called n-grams, with windows that glide over the sequence of notes. The resulting strings can be indexed with a normal text search engine. This approach can also be extended to polyphonic music, where more than one note can be sounded at any one time (Doraisamy and Rüger, 2003).

0.2.2 Automated annotation

Two of the factors limiting the uptake of digital libraries for multimedia are the scarcity and the expense of metadata for digital media. Flickr[4], a popular photo sharing site, lets users upload, organise and annotate their own photographs with tags. In order to search images in Flickr, little more than user tags are available with the effect that many photographs are difficult or impossible to find. The same is true for the video sharing site YouTube[5]. At the other end of the spectrum are commercial sites such as the digital multimedia store iTunes[6], which sells music, movies, TV shows, audio-books, podcasts and games. They tend to have sufficiently many annotations as the commercial nature of iTunes makes it viable to supply metadata to the required level of granularity. While personal photographs and videos do not come with much metadata except for the data that the camera provides (time-stamp and technical data such as aperture, exposure, sensitivity and focal length), a whole class of surveillance data carries even less incentive to create metadata manually: CCTV recordings, satellite images, audio recordings in the sea and other sensor data. The absence of labels and metadata is a real barrier for complex and high-level queries such as "what did the person with a red jumper look like who exited the car park during the last 6 hours in a black Volvo at high speed".

One way to generate useful tags and metadata for multimedia objects is to involve a community of people who carry out the tagging collaboratively. This process is also called folksonomy, social indexing or social tagging. Del.icio.us[7] is a social bookmarking system and a good example for folksonomies. Similarly, the ability of Flickr to annotate images of other people falls also into this category. Von Ahn and Dabbish (2004) have invented a computer game that provides an incentive (competition and points) for people to label randomly selected images. All these approaches tap into "human computing power" for a good cause: the structuring and labelling of multimedia objects. Research in this area is still in the beginning, and it is by no way clear how to best harness the social power of collaborative tagging to improve metadata for, and access to, digital museums and

[4] http://flickr.com
[5] http://www.youtube.com
[6] http://www.apple.com/itunes
[7] http://del.icio.us

libraries.

Another way to bridge the semantic gap (see Section 0.3) is to try to assign simple words automatically to images solely based on their pixels. Methods attempting this task include dedicated machine vision models for particular words such as "people" or "aeroplane". These individual models for each of the words can quickly become very detailed and elaborate: Thomas Huang of the University of Illinois at Urbana Champaign once joked during his keynote speech at CIVR 2002 that in order to enable a system to annotate 1,000 words automatically, it was merely a case of supervising 1,000 corresponding PhD projects!

Automated annotation can be formulated in more general terms of machine translation as seen in Figure 4. The basic idea is to first dissect images into blobs of similar colour and then use these blobs as "words" of a visual vocabulary. Given a training set of annotated images a correlation between certain words and certain blobs can then be established in a similar way to correlations between corresponding words of two different languages using a parallel corpus (for example, the official records of the Canadian Parliament in French and English). Duygulu et al (2002) created the first successful automated annotation mechanisms based on this idea.

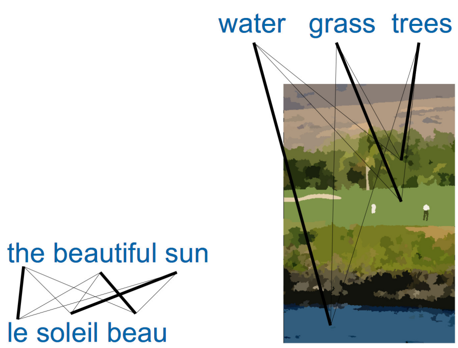


Figure 4: Automated annotation as machine translation problem

However, the most popular and successful *generic* approaches are based on classification techniques. This normally requires a large training set of images that have annotations from which one can extract features and correlate these with the existing annotations of the training set. For example, images with tigers will have orange-black stripes and often green patches from surrounding vegetation, and their existence in an unseen image can in turn bring about the annotation "tiger". As with any machine learning method, it is important to work with a large set of training examples. Figure 5 shows randomly selected, royalty free images from the Corel's Gallery 380,000 product that were annotated with *sunset* (top) and *city* (bottom). Each of these images can have multiple annotations: there are pictures that are annotated with *both* sunset and city, and possibly other terms.

Automated algorithms build a model for the commonalities in the features of images, which can later be used for retrieval. One of the simplest machine learning algorithms is the Naïve Bayes

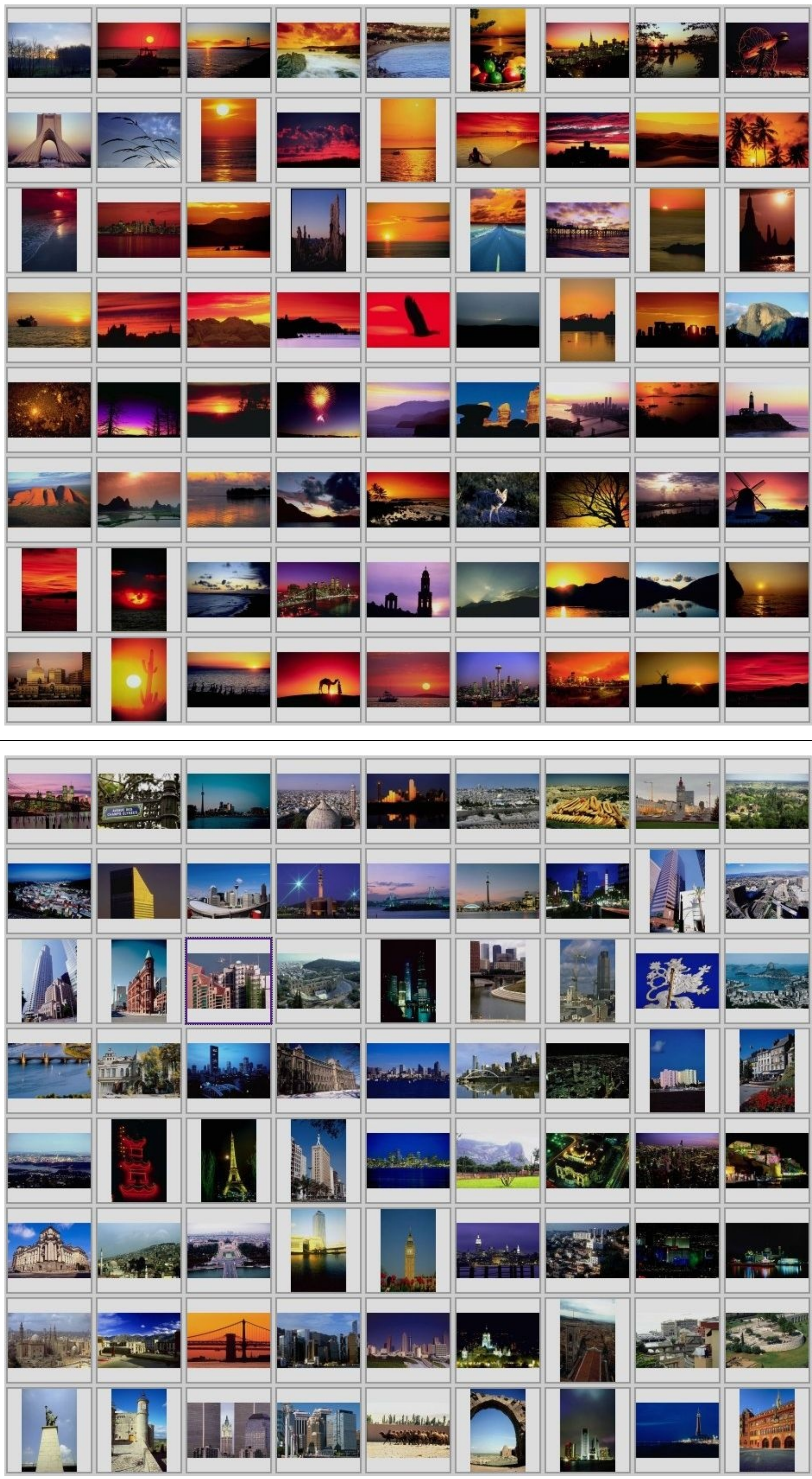


Figure 5: Machine learning training samples for *sunset* (top) and *city* images (bottom)

formula,

$$P(w|i) = \frac{P(w,i)}{P(i)} = \frac{\sum_j P(w,i|j)P(j)}{\sum_j P(i|j)P(j)}$$

$$= \frac{\sum_j P(i|w,j)P(w|j)P(j)}{\sum_j \sum_w P(i|w,j)P(w|j)P(j)},$$

where j are training images, w are word annotations and $P(w|i)$ is the probability of a word w given an (unseen) image i. The probability $P(w,j)$ that word w is used to annotate image j can be estimated from an empirical distribution of annotations in the training data.

Figure 6 shows an unseen image i for which the five words with the highest probabilities $p(w|i)$ according to above Naïve Bayes classification are all sensible and useful.

Figure 6: Automated annotation results in *water*, *buildings*, *city*, *sunset* and *aerial*

Yavlinsky et al (2005) built models based on a similar idea for which the model for keywords appearance is derived from non-parametric density estimators with specialised kernels that utilise the Earth mover's distance. The assumption is that these kernels reflect the nature of the underlying features well. Yavlinsky built a corresponding search engine behold[8], where one could search for Flickr images using these detected terms. These algorithms all make errors as one can expect from fully automated systems. Figure 7 shows screenshots from an early version of behold. Clearly, not all words are predicted correctly, and the ugly examples from this figure might motivate to study methods that use external knowledge, for example, that stairs and icebergs normally do not go together.

Today, Makadia et al's recent (2008) work on the nearest neighbour label transfer provide a baseline for automatic image annotation using global low-level features and a straightforward label transfer from the 5 nearest neighbours. This approach is likely to work very well if enough images are available in a labelled set that are very close to the unlabelled application set. This may be the case, for example, in museums where images of groups of objects are taken in a batch fashion with the same lighting and background and only some of the objects in the group have received manual labels. Liu et al (2009a) also use label transfer, albeit in a slightly different setting since they aim to segment and recognise scenes rather than assign global classification labels.

Automated annotation from pixels faces criticism not only owing to its current inability to model a large and useful vocabulary with high accuracy. Enser and Sandom (2002, 2003) argue that some

[8]http://www.behold.cc

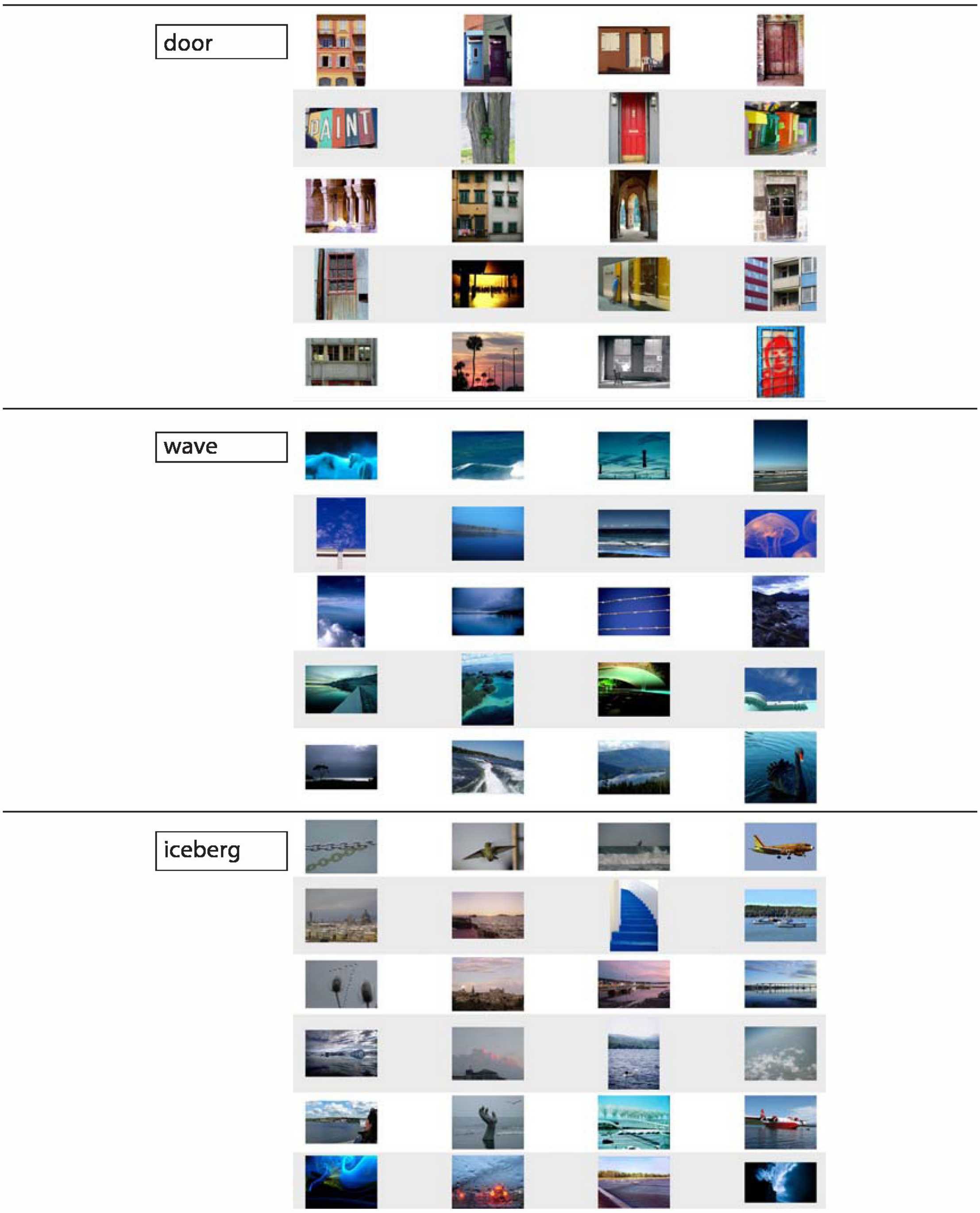


Figure 7: The good, the bad and the ugly: three examples for automated annotation

of the vital information for significance and content of images *has* to come from metadata: it is virtually impossible to, eg, compute the date or location of an image from its pixels. A real-world image query such as "Stirling Moss winning Kentish 100 Trophy at Brands Hatch, 30 August 1968" cannot be answered without metadata. They argue that pixel-based algorithms will never be able to compute *significance* of images such as "first public engagement of Prince Charles as a boy" or "the first ordination of a woman as bishop". Their UK-funded arts and humanities research project "Bridging the Semantic Gap in Visual Information Retrieval" (Hare et al, 2006; Enser and Sandom, 2003) brought a new understanding about the role of the semantic gap in visual image retrieval.

Owing to these observations and also owing to their relatively large error rates, automated annotation methods seem to be more suitable in the context of browsing or in conjunction with other search methods. For example, if you want to "find shots of the front of the White House in the daytime with the fountain running"[9], then a query-by-example search in a large database may be solved quicker and better by emphasising those shots that were classified as "vegetation", "outside", "building" etc — even though the individual classification may be wrong in a significant proportion of cases.

There is a host of research that supports the bridging of the semantic gap via automated annotation. Hare and Lewis (2004) use salient interest points and the concept of scale to the selection of salient regions in an image to describe the image characteristics in that region; they then extended this work (2005) to model visual terms from a training set that can then be used to annotate unseen images. Magalhães and Rüger (2006) developed a clustering method that is more computationally efficient than the currently very effective method of non-parametric density estimation, which they later (2007) integrated into a unique multimedia indexing model for heterogeneous data. Torralba and Oliva (2003) obtained relatively good results with simple scene-level statistics, while others deploy more complex models: Jeon et al (2003) and Lavrenko et al (2003) studied cross-lingual information retrieval models, while Metzler and Manmatha (2004) set up inference networks that connect image segments with words. Blei and Jordan (2003) carry out probabilistic modelling with latent Dirichlet allocation, while Feng et al (2004) use Bernoulli distributions.

Machine learning methods for classification and annotation are not limited to images at all. For example, one can extract motion vectors from MPEG-encoded videos and use these to classify a video shot independently into categories such as object motion from left to right, zoom in, tilt, roll, dolly in and out, truck left and right, pedestal up and down, crane boom, swing boom etc. In contrast to the above classification tasks, the extracted motion vector features are much more closely correlated to the ensuing motion label than image features are to text labels, and the corresponding learning task should be much simpler a consequence.

The application area for classification can be rather diverse: Baillie and Jose (2004) use audio analysis of the crowd response in a football game to detect important events in the match; Cavallaro and Ebrahimi (2004) propose an interaction mechanism between the semantic and the region partitions, which allows to detect multiple simultaneous objects in videos.

On a higher level, Salway and Graham (2003) developed a method to extract information about emotions of characters in films and suggested that this information can help describe higher levels of multimedia semantics relating to narrative structures. Salway et al (2005) contributed to the analysis and description of semantic video content by investigating what actions are important in films.

Musical genre classification can be carried out on extracted audio-features that represent a performance by its statistics of pitch content, rhythmic structure and timbre texture (Tzanetakis

[9]Topic 124 of TRECVid 2003, see http://www-nlpir.nist.gov/projects/tv2003

and Cook, 2002): timbre texture features are normally computed using short-time Fourier transform and Mel-frequency cepstral coefficients that also play a vital role in speech recognition; the rhythmic structure of music can be explored using discrete wavelet transforms that have a different time resolution for different frequencies; pitch detection, especially in polyphonic music, is more intricate and requires more elaborate algorithms. For details, see the work of Tolonen and Karjalainen (2000). Tzanetakis and Cook (2002) report correct classification rates of between 40% (rock) and 75% (jazz) in their experiments with 10 different genres.

0.2.3 Content-based retrieval

Content-based retrieval uses characteristics of the multimedia objects themselves, ie, their content to search and find multimedia. Its main application is to find multimedia by examples, ie, when the query consists not of words but of a similar example instance.

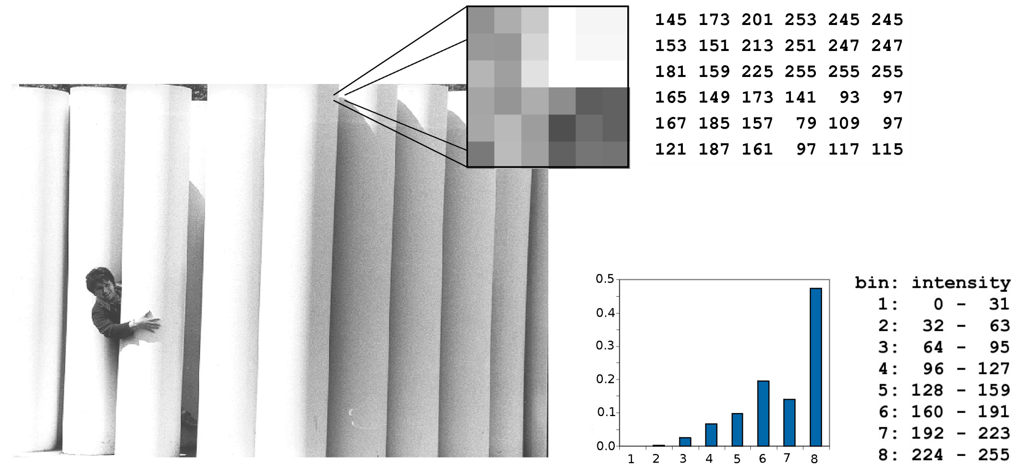


Figure 8: Millions of pixels with intensity values and the corresponding intensity histogram

One of the difficulties of matching multimedia is that the parts the media are made from are not necessarily semantic units. Another difficulty comes about by the sheer amount of data with little apparent structure. Look at the black and white photograph of Figure 8, for example. It literally consists of millions of pixels, and each of the pixels encodes an intensity (one number between 0=black and 255=white) or a colour (three numbers for the red, green and blue colour channel, say). One of the prime tasks in multimedia retrieval is to make sense out of this sea of numbers.

The key here is to condense the sheer amount of numbers into meaningful pieces of information, which we call *features*. One trivial example is to compute an intensity histogram, ie, count which proportion of the pixels falls into which intensity ranges. In Figure 8 I have chosen 8 ranges, and the histogram of 8 numbers conveys a rough distribution of brightness in the image.

Figure 9 shows the main principle of *query-by-example*Query-by-Example; in this case, the query is the image of an ice-bear on the left. This query image will have a representation as a certain point (o) in feature space. In the same way, every single image in the database has its own representation (x) in the same space. The images, whose representations are closest to the representation of the query are ranked top by this process. The two key elements really are features and distances. Our choice of feature space and how to compute distances has a vital impact on how well visual search by example works.

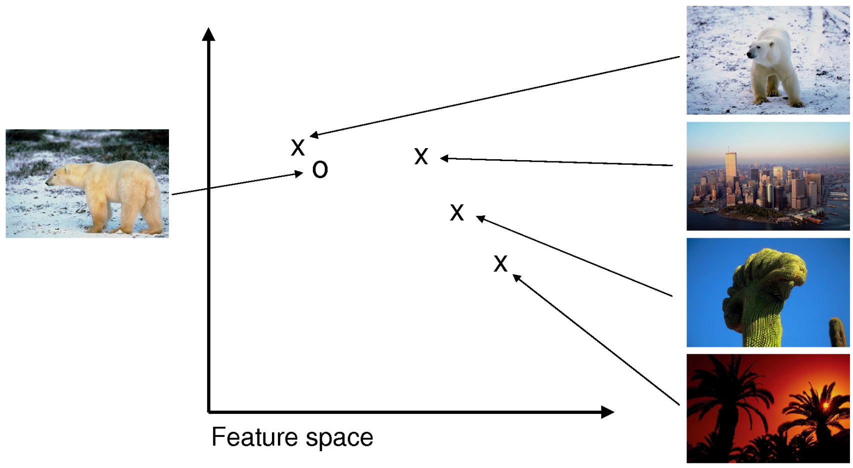


Figure 9: Features and distances

Features and distances are a vital part of content-based retrieval and so is the ability to efficiently find nearest neighbours in high-dimensional spaces. Lew et al (2006) and Datta et al (2008) have published overview articles on content-based retrieval, and Rüger (2010) treats content-based retrieval including features and distances commonly used in depth.

The architecture presented here is a typical albeit basic one; there are many variations and some radically different approaches that have been published in the past. A whole research field has gathered around the area of video and image retrieval as exemplified by the annual International ACM Conferences on Video and Image Retrieval (CIVR), Multimedia (ACM MM) and Multimedia Information Retrieval (MIR) and the TREC video evaluation workshop TRECVid; there is another research field around music retrieval, see the annual International Conference on Music Information Retrieval (ISMIR).

0.2.4 Fingerprinting

Multimedia fingerprints are unique indices in a multimedia database. They are computed from the contents of the multimedia objects, are small, allow the fast, reliable and *unique* location of the database record and are robust against degradation or deliberate change of the multimedia document that do not alter their human perception. Audio fingerprints of music tracks are expected to distinguish even between different performances of the same song by the same artist at perhaps different concerts or studios.

Interesting applications include services that allow broadcast monitoring companies to identify what was played, so that royalties are fairly distributed or programmes and advertisements verified. Other applications uncover copyright violation or, for example, provide a service that allows you to locate the meta-data such as title, artist and date of performance from snippets recorded on a (noisy) mobile phone.

Cano et al (2005) review some audio fingerprinting methods and Seo et al (2004) proposes an image fingerprinting technique.

0.3 Challenges of automated visual indexing

There are a number of open issues with the content-based retrieval approach in multimedia. On a perceptual level, those low-level features do not necessarily correlate with any high-level meaning the images might have. This problem is known as the *semantic gap*: imagine a scene in which Bobby Moore, the captain of the English National Football team in 1966, receives the world cup trophy from Queen Elizabeth II; there is no obvious correlation between low-level colour, shape and texture descriptors and the high-level meaning of victory and triumph (or defeat and misery if you happened to support the West German team). Some of the computer vision methods go towards the bridging of the semantic gap, for example the ability to assign simple concrete labels to image parts such as “grass”, “sky”, “people”, “plates”. A consequent use of an ontology could explain the presence of higher-level concepts such as “barbecue” in terms of the simpler labels.

Even if the semantic gap could be bridged, there is still another challenge, namely *polysemy*: images usually convey a multitude of meanings so that the query-by-example approach is bound to under-specify the real information need. Users who submit an image such as the one in Figure 8 could have a dozen different information needs in mind: “find other images with the same person”, “find images of the same art scene”, “find other bright art sculptures”, “find images with gradual shadow transitions”, ... It is these different interpretations that make further user feedback so important.

User feedback can change the weights of features in content-based retrieval scenarios; these weights represent the plasticity of the retrieval system. Hence, putting the user in the loop and designing a human-computer interaction that utilises the user's feedback has been one of the main approaches to tackle these perceptual issues. Amongst other methods there are those that seek to reformulate the query (Ishikawa et al, 1998) or those that weight the various features differently depending on the user's feedback. Weight adaptation methods include cluster analysis of the images (Wood et al, 1998); transposed files for feature selection (Squire et al, 2000); Bayesian network learning (Cox et al, 2000); statistical analysis of the feature distributions of relevant images and variance analysis (Rui et al, 1998); and analytic global optimisation (Heesch and Rüger, 2003). Some approaches give the presentation and placement of images on screen much consideration to indicate similarity of images amongst themselves (Santini and Jain, 2000; Rodden et al, 1999) or with respect to a visual query (Heesch and Rüger, 2003).

On a practical level, the multitude of features assigned to images poses a *fusion problem*; how to combine possibly conflicting evidence of two images' similarity? There are many approaches to carry out fusion, some based on labelled training data and some based on user feedback for the current query (Aslam and Montague, 2001; Bartell et al, 1994; Shaw and Fox, 1994; Yavlinsky et al, 2004).

There is a *responsiveness problem*, too, in that the naïve comparison of query feature vectors to the database feature vectors requires a linear scan through the database. Although the scan is eminently scalable, the practicalities of doing this operation can mean an undesirable response time in the order of seconds rather than the 100 milli-seconds that can be achieved by text search engines. The problem is that high-dimensional tree structures tend to collapse to linear scans above a certain dimensionality (Weber et al, 1998). As a consequence, some approaches for fast nearest-neighbour search use compression techniques to speed up the disk access of linear scan as in (Weber et al, 1998) using VA-files; or they approximate the search (Nene and Nayar, 1997; Beis and Lowe, 1997); decompose the features componentwise (de Vries et al, 2002; Aggarwal and Yu, 2000) saving access to unnecessary components; or deploy a combination of these (Müller and Henrich, 2004;

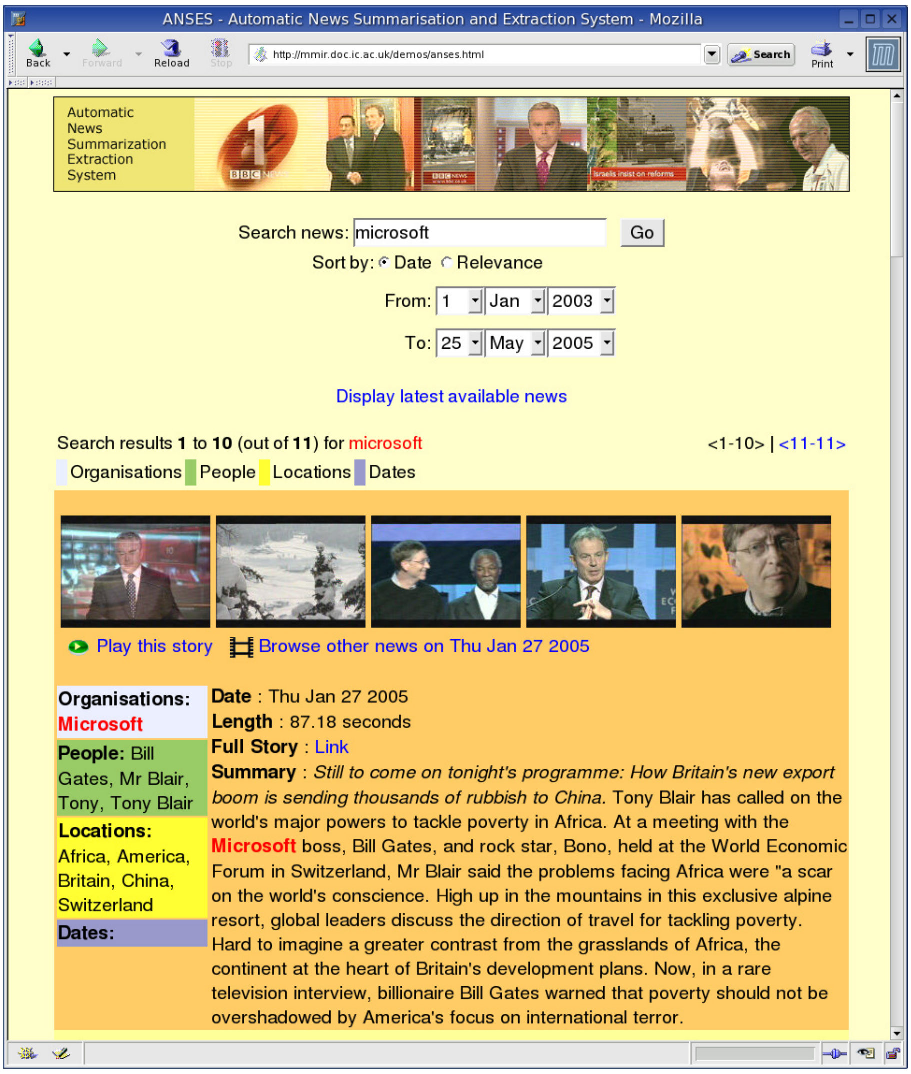


Figure 10: News search engine interface

Howarth and Rüger, 2005c).

0.4 Added Services

0.4.1 Video summaries

Even if the challenges of the previous section were all solved and if the automated methods of Section 0.2 enabled a retrieval process with high precision (proportion of the retrieved items that are relevant) and high recall (proportion of the relevant items that are retrieved) it would still be vital to present the retrieval results in a way so that the users can quickly decide to which degree those items are relevant to them.

Images are most naturally displayed as thumbnails, and their relevance can quickly be judged by users. Presenting and summarising videos is a bit more involved. The main metaphor used for this is that of a *storyboard* that contains *keyframes* with some text about the video. Several systems exist that summarise news stories in this way, most notably Informedia (Christel et al, 1999) and Físchlár (Smeaton et al, 2004). The Informedia system devotes much effort to added services such as face recognition and speaker voice identification allowing retrieval of the appearance of known people. Informedia also provides alternative modes of presentation, eg, through film skims or by assembling 'collages' of images, text and other information (eg, maps) sourced via references from the text (Christel and Warmack, 2001). Físchlár's added value lies in the ability to personalise the content (with the user expressing like or dislike of stories) and in assembling lists of related stories and recommendations.

Our very own TV news search engine ANSES (Pickering et al, 2003; Pickering, 2004) records the main BBC evening news along with the sub-titles, indexes them, breaks the video stream into shots (defined as those video sequences that are generated during a continuous operation of the camera), extracts one key-frame per shot, automatically glues shots together to form news stories based on an overlap in vocabulary in the sub-titles of adjacent shots (using lexical chains), and assembles a story-board for each story. Stories can be browsed or retrieved via text searches. Fig 10 shows the interface of ANSES. We use the natural language toolset GATE (Cunningham, 2002) for automated discovery of organisations, people, places and dates; displaying these prominently as part of a storyboard as in Figure 10 provides an instant indication of what the news story is about. ANSES also displays a short automated textual extraction summary, again using lexical chains to identify the most salient sentences. These summaries are never as informative as hand-made ones, but users of the system have found them crucial for judging whether or not they are interested in a particular returned search result.

Dissecting the video stream into shots and associating one keyframe along with text from subtitles to each shot has another advantage: A video collection can essentially be treated as an image collection, where each, possibly annotated image acts as entry point into the video.

0.4.2 New Paradigms in Information Visualisation

The last decade has witnessed an explosion in interest in the field of information visualisation, (Hemmje et al, 1994; Ankerst et al, 1996; Card, 1996; Shneiderman et al, 2000; Börner, 2000). Here we present three new visualisation paradigms, based on our earlier design studies (Au et al, 2000; Carey et al, 2003). These techniques all revolve around a representation of documents in the form of bag-of-words vectors, which can be clustered to form groups. We use a variant of the buckshot clustering algorithm for this. Basically, the top, say, 100 documents that were returned from a query are clustered via hierarchical clustering to initialise document centroids for k-means clustering that puts all documents returned by a query into groups. Another common element of our visualisations is the notion of *keywords* that are specific to the returned set of documents. The keywords are computed using a simple statistic; for details see (Carey et al, 2003). The new methods are:

Sammon Cluster View. This paradigm uses a Sammon map to generate a two dimensional screen location from a many-dimensional vector representing a cluster centroid. This map is computed using an iterative gradient search (Sammon, 1969) while attempting to preserve the pairwise distances between the cluster centres. Clusters are thus arranged so that their mutual distances are indicative of their relationship. The idea is to create a visual landscape for navigation. Fig 11 shows an example of such an interface. The display has three panels, a scrolling table panel to the left, a graphic panel in the middle and a scrolling text panel to the right that contains the

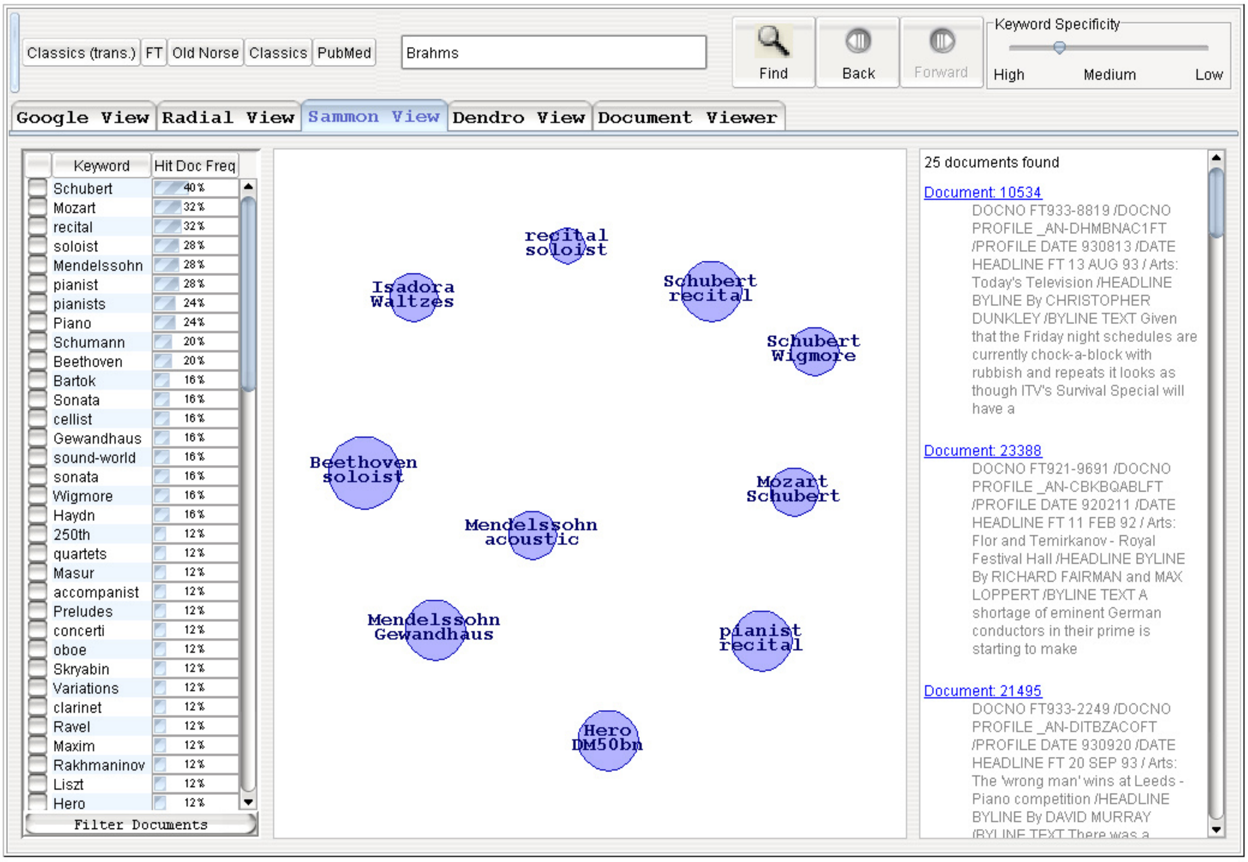


Figure 11: Sammon map for cluster-guided search

traditional list of returned documents as hotlinks and snippets. In the graphic panel each cluster is represented by a circle and is labelled with its two most frequent keywords. The radius of the circle represents the cluster size. The distance between any two circles in the graphic panel is an indication of the similarity of their respective clusters - the nearer the clusters, the more likely the documents contained within will be similar. When the mouse passes over the cluster circle a tool-tip box in the form of a pop-up menu appears that allows the user to select clusters and *drill down*, ie, re-cluster and re-display only the documents in the selected clusters. The back button undoes this process and climbs up the hierarchy (*drill up*). The table of keywords includes box fields that can be selected. At the bottom of the table is a filter button that makes the scrolling text window display only the hot-links and snippets from documents that contain the selected keywords.

Dendro Map Visualisation. The Dendro Map visualisation represents documents as leaf nodes of a binary tree that is output by the buckshot clustering algorithm. With its plane-spanning property and progressive shortening of branches towards the periphery, the Dendro Map mimics the result of a non-Euclidean transformation of the plane as used in hyperbolic maps without suffering from their computational load. Owing to spatial constraints, the visualisation depth is confined to five levels of the hierarchy with nodes of the lowest level representing either documents or subclusters. Different colours facilitate visual discrimination between individual documents and clusters. Each lowest level node is labelled with the most frequent keyword of the subcluster or document. This

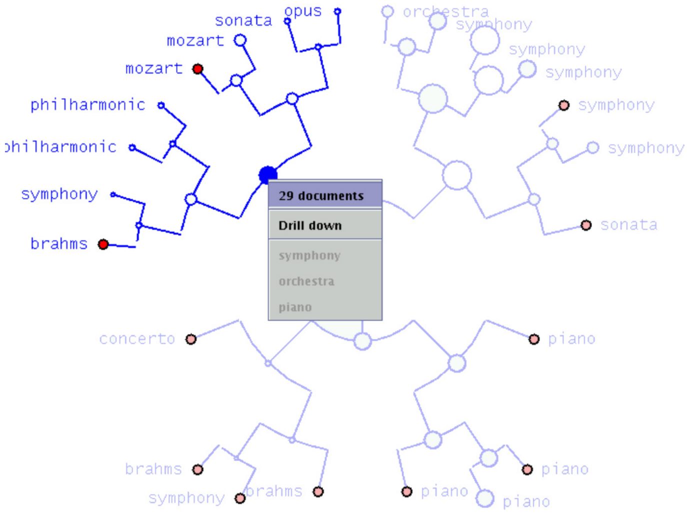


Figure 12: Dendro Map - A plane-spanning binary tree (query “Beethoven”)

forms a key component of the Dendro Map as it gives the user the cues needed for navigating through the tree. As the user moves the mouse pointer over an internal node, the internal nodes and branches of the associated subcluster change colour from light blue to dark blue while the leaf nodes, ie, document representations, turn bright red. As in the Sammon Map, a tool-tip window provides additional information about the cluster and can be used to display a table with a list of keywords associated with the cluster. The user may drill down on any internal node. The selected node will as a result replace the current root node at the center and the entire display is re-organized around the new root. The multi-level approach of the Dendro Map allows the user to gain a quick overview over the document collection and to identify promising subsets.

Radial Interactive Visualisation. Radial (Figure 13) is similar to VIBE (Korfhage, 1991), to Radviz (Hoffman et al, 1999) and to Lyberworld (Hemmje et al, 1994). It places the keyword nodes round a circle, and the position of the document dots in the middle depend on the force of invisible springs connecting them to keyword nodes: the more relevant a keyword for a particular document, the stronger its spring pulls on the document. Hence, we make direct use of the bag-of-words representation without explicit clustering. Initially, the twelve highest ranking keywords are displayed in a circle. The interface lets the user move the keywords, and the corresponding documents follow this movement. This allows the user to manually cluster the documents based on the keywords they are interested in. As the mouse passes over the documents, a bubble displays a descriptive piece of text. The location of document dots is not unique owing to dimensionality reduction, and there may be many reasons for a document to have a particular position. To mitigate this ambiguity in Radial the user can click on a document dot, and the keywords that affect the location of document are highlighted. A choice of keywords used in the display can be exercised by clicking on two visible lists of words. Zoom buttons allow the degree of projection to be increased or reduced so as to distinguish between documents around the edges of the display or at the centre. The Radial visualisation appears to be a good interactive tool to structure the document set according to one's own preferences by shifting keywords around in the display.

Unified Approach. The integration of the paradigms into one application offers the possibility of browsing the same result set in several different ways simultaneously. The cluster-based visualisations give a broader overall picture of the result, while the Radial visualisation allows the user

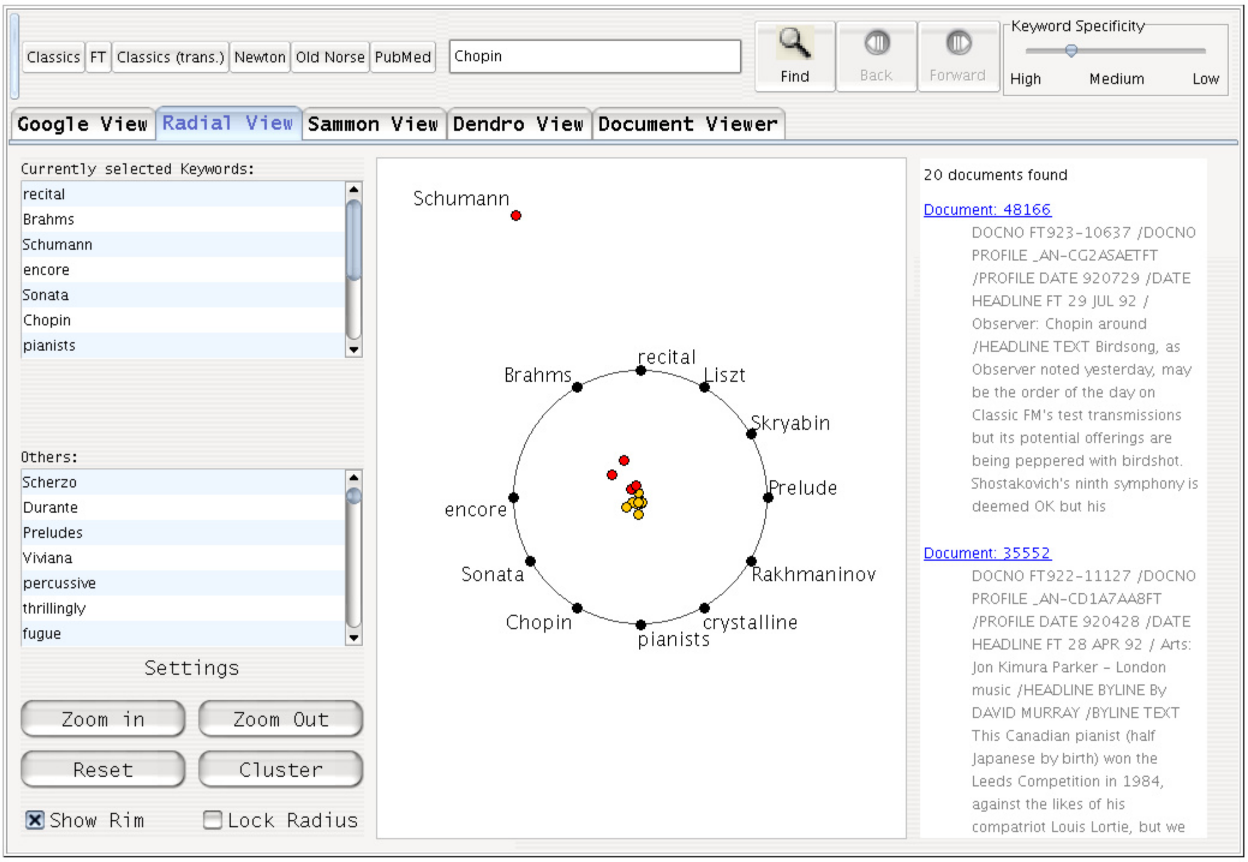
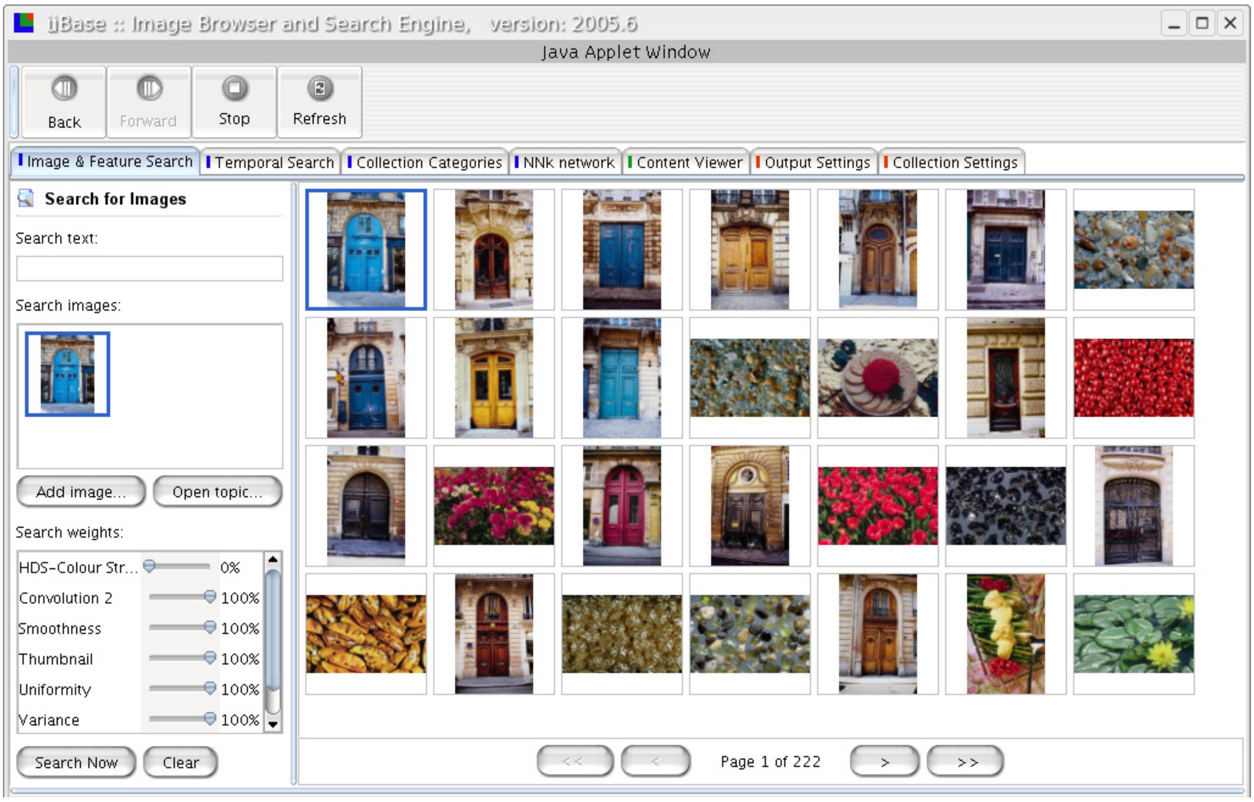
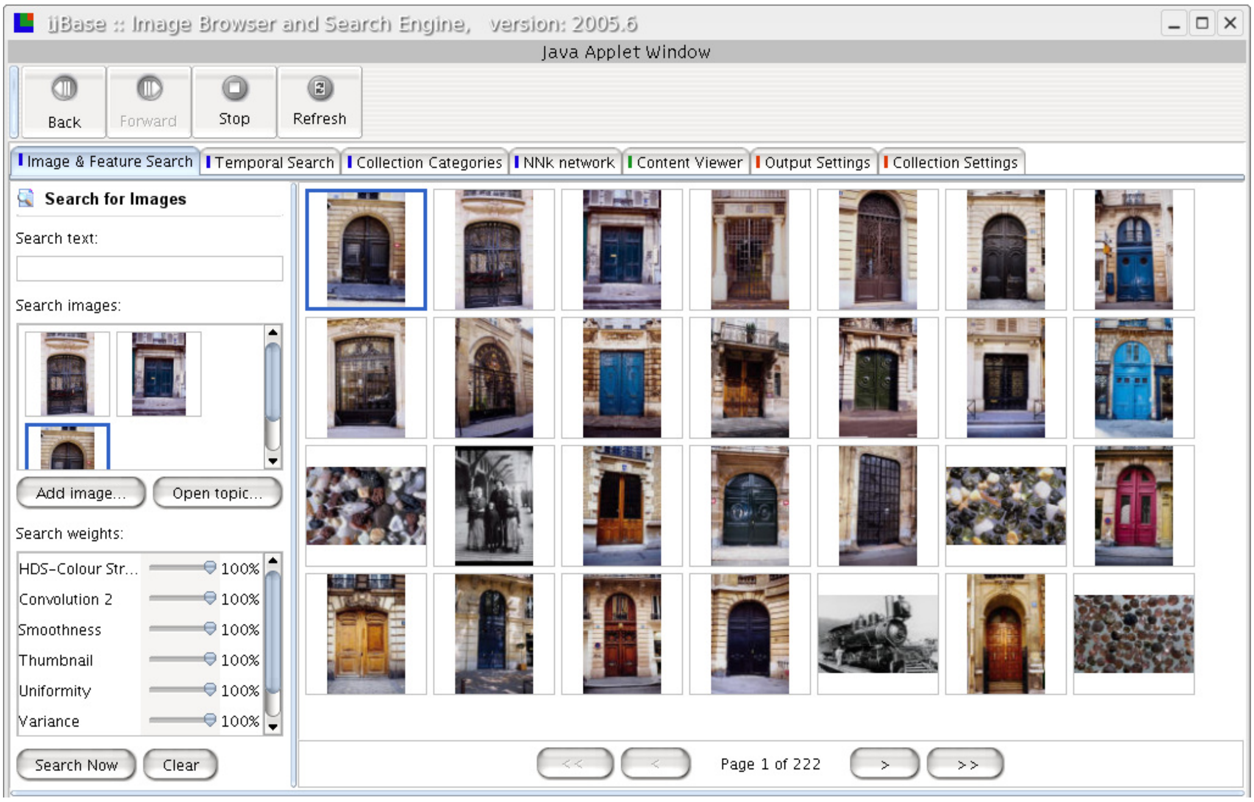


Figure 13: Radial Visualisation



(a) Query by example (left panel) with initial results in the right panel



(b) A new query made of three images from (a) results in many more dark-door images

Figure 14: Visual search for images of dark doors starting with a bright-door example

to focus on subsets of keywords. Also, as the clusters are approximations that highlight particular keywords, it may be useful to return to the Radial visualisation and examine the effect of these keywords upon the whole document set. The Radial visualisation will perhaps be more fruitful if the initial keywords match the user's area of interest. The Sammon Map will let the user dissect search sets and re-cluster subsets, gradually homing in on target sets. This interface was developed within the joint NSF-EC project CHLT (http://www.chlt.org); it was evaluated from a human-computer-interaction point of view with encouraging results (Chawda et al, 2005) and has proven useful in real-world multi-lingual scholarly collections (Rydberg-Cox et al, 2004).

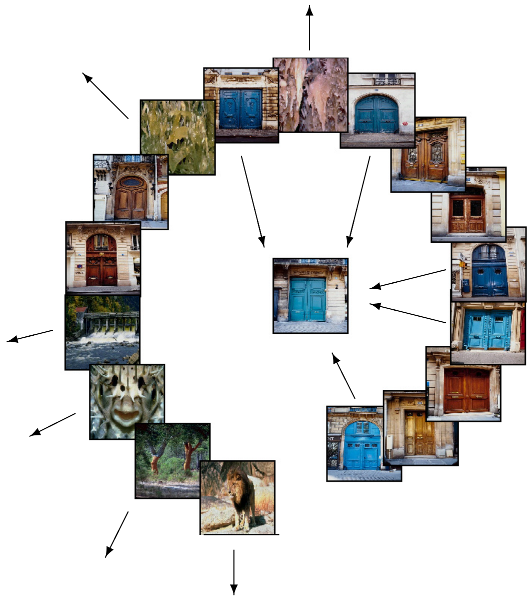


Figure 15: A relevance feedback model

0.4.3 Visual search and relevance feedback

The visual query-by-example paradigm discussed in Section 0.3 gives rise to relatively straightforward interfaces; an image is dragged into a query box, or, eg, specified via a URL, and the best matching images are displayed in a ranked list to be inspected by the user, see Fig 14(a). A natural extension of such an interface is to offer the selection of relevant results as new query elements. This type of relevance feedback, a.k.a. *query point moving*, is shown in Fig 14(b).

One other main type of relevance feedback, *weight space movement*, assumes that the relative weight of the multitude of features that one can assign to images (eg, structured meta-data fields such as author, creation date and location; low-level visual features such as colour, shape, structure and texture; free-form text) can be learned from user feedback. Of the methods mentioned in Section 0.3 our group chose analytic weight updating as this has a very small execution time. The idea is that users can specify the degree to which a returned image is relevant to their information needs. This is done by having a visual representation; the returned images are listed in a spiral, and the distance of an image to the centre of the screen is a measure of the relevance that the search engine assigns to a specific image. Users can now move the images around with the mouse or place them in the centre with a left mouse click and far away with a right click. Fig 15 shows this relevance feedback model. We evaluated the effectiveness of negative feedback, positive feedback and query point moving, and found that combining the latter two yields the biggest improvement in terms of mean average precision (Heesch and Rüger, 2003).

A new and relatively unexplored area of relevance feedback is the exploitation of social context information. By looking not only at the behaviour and attributes of the user, but also his past interactions and also the interactions of people he has some form of social connection with could yield useful information when determining whether search results are relevant or not. Browsing systems could recommend data items based on the actions of a social network instead of just a single user, using more data to yield better results.

The use of such social information is also becoming important for multimedia meta data generation, particular in the area of folksonomies where the feedback of users actively produces the terms and taxonomies used to describe the media in the system instead of using a predetermined, prescribed dictionary (Voss, 2007). This can be seen being effectively used in online multimedia systems such as Flickr (http://www.flickr.com) and del.icio.us (http://del.icio.us).

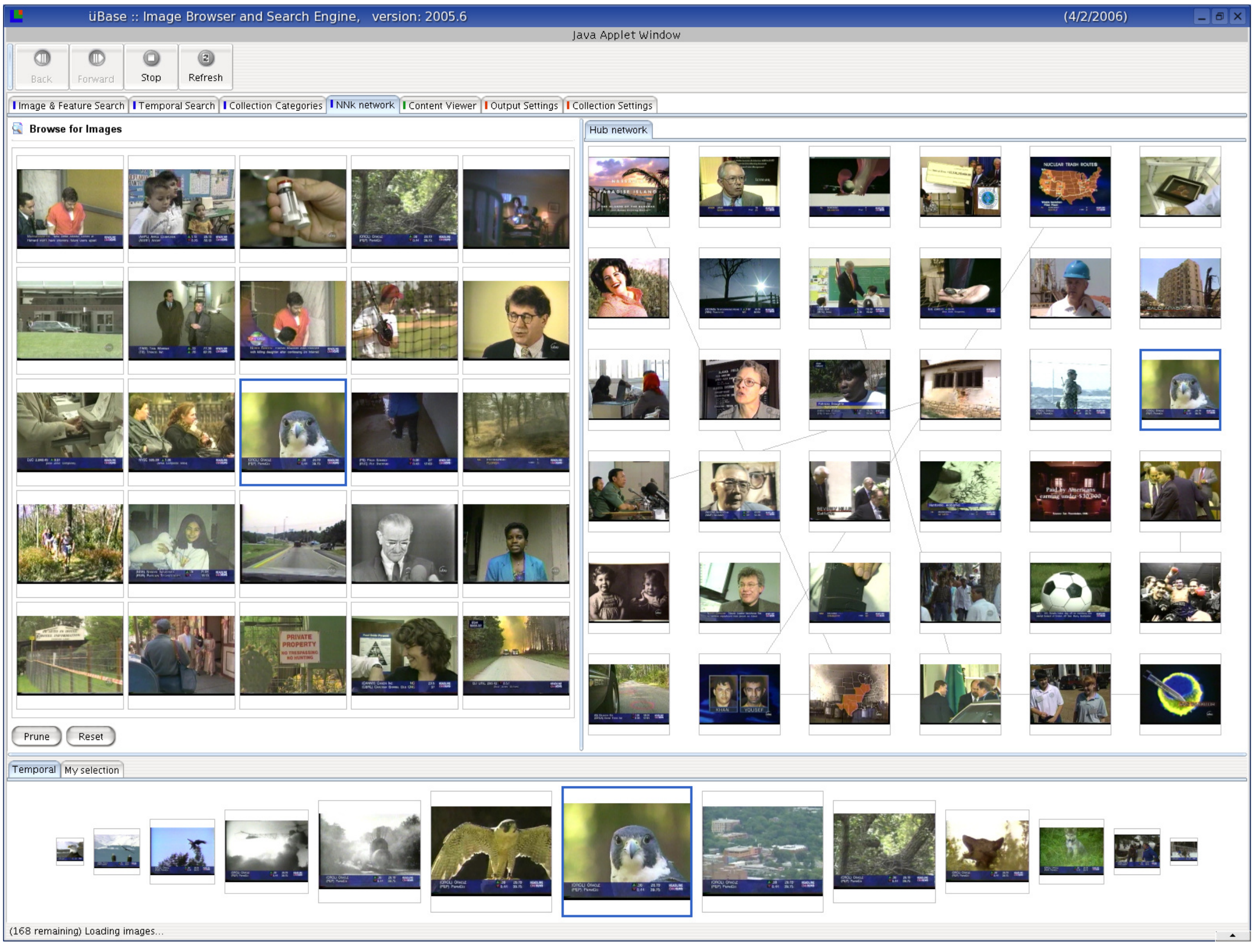
0.5 Browsing: lateral and geo-temporal

The idea of representing text documents in a nearest-neighbour network was first presented by Croft and Parenty (1985), albeit, as an internal representation of the relationships between documents and terms, not for browsing. Document networks for interactive browsing were identified by Cox (1992 and 1995). Attempts to introduce the idea of browsing into content-based image retrieval include Campbell's work 2000; his ostensive model retains the basic mode of query based retrieval but in addition allows browsing through a dynamically created local tree structure. Santini and Jain's *El niño* system 2000 is another attempt to combine query-based search with browsing. The system tries to display configurations of images in feature space such that the mutual distances between images are preserved as well as possible. Feedback is given in the same spirit as in Fig 15 by manually forming clusters of images that appear similar to the user. This in turn results in an altered configuration with potentially new images being displayed.

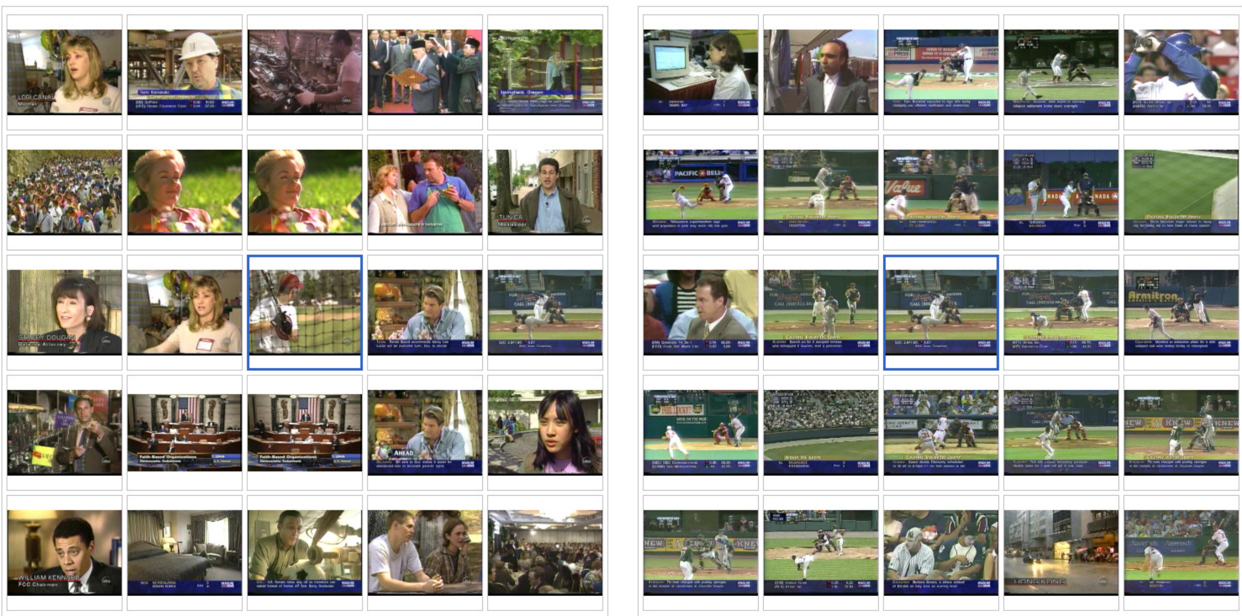
Other network structures that have increasingly been used for information visualisation and browsing are Pathfinder networks (Dearholt and Schvaneveldt, 1990). They are constructed by removing redundant edges from a potentially much more complex network. Fowler et al (1992) use Pathfinder networks to structure the relationships between terms from document abstracts, between document terms and between entire documents. The user interface supports access to the browsing structure through prominently marked high-connectivity nodes.

Our group (Heesch and Rüger, 2004) determines the nearest neighbour for the image under consideration (which we call the *focal* image) for *every* combination of features. This results in a set of what we call *lateral neighbours*. By calculating the lateral neighbours of all database images, we generate a network that lends itself to browsing. Lateral neighbours share some properties of the focal image, but not necessarily all. For example, a lateral neighbour may share text annotations with the focal image, but no visual similarity with it at all, or it may have a very similar colour distribution, but no structural similarity, or it may be similar in all features except shape, etc. As a consequence, lateral neighbours are deemed to expose the polysemy of the focal image. Hence, when they are presented, the user may then follow one of them by making it the focal image and explore its lateral neighbours in turn. The user interaction is immediate, since the underlying network was computed offline.

We provide the user with entry points into the database by computing a representative set of images from the collection. We cluster high-connectivity nodes and their neighbours up to a

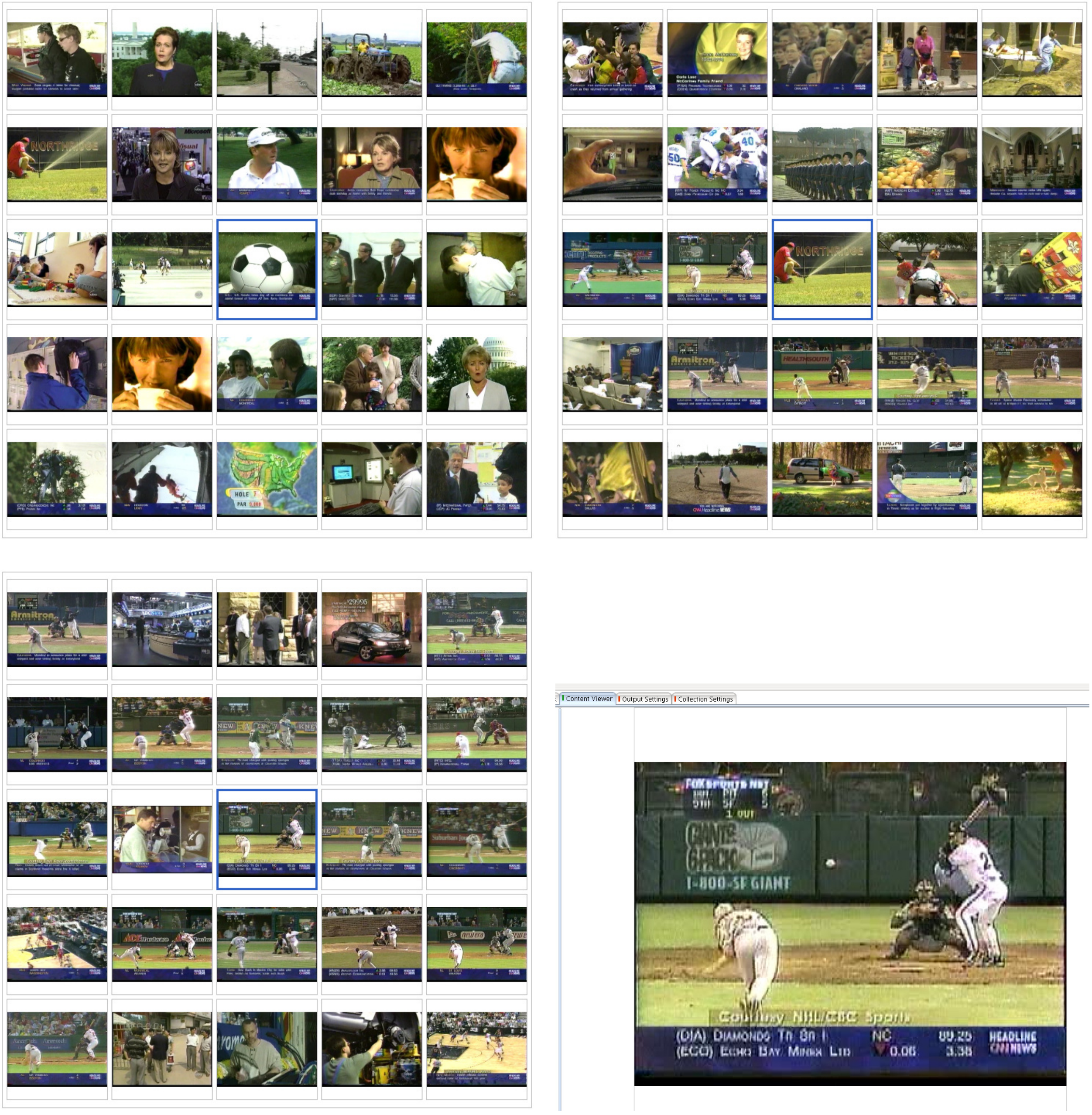


(a) Initial visual summary of the database (right panel) from which the user chooses the falcon, its nearest lateral neighbours are then displayed in the left panel.



(b) Clicking on any image will make it the centre of the nearest neighbours panel and display is associated lateral neighbours around it.

Figure 16: Lateral browsing for an image "from behind the pitcher in a baseball game..."



Starting with the football image (upper left) from the database overview, one of its lateral neighbours is an image of a lawn with a sprinkler; when this is made the focal image (upper right) there are already images from baseball scenes. Clicking on one of them (lower left) reveals that there are more of this kind; they can be enlarged and the corresponding video played in the "viewer tab" (lower right).

Figure 17: Alternative ways to browse for images "from behind the pitcher . . ."

certain depth using the Markov chain clustering algorithm (van Dongen, 2000), which has robust convergence properties and allows one to specify the granularity of the clustering. The clustering result can be seen as a image database summary that shows highly-connected nodes with far-reaching connections. The right panel of Fig 16(a) is such a summary for our TRECVid (2003) database. The user may select any of these images as an entry point into the network. Clicking on an image moves it into the centre around which the lateral neighbours are displayed, see the nearest-neighbour panel on the left side of fig 16(a). If the size of the lateral-neighbour set is above a certain threshold the actual number of images displayed is reduced to the most salient ones.

If a user wanted to find "video shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at" (TRECVid, 2003, topic 102) then they might explore the database in Fig 16 by clicking on the falcon image. The hope is that the colour of a baseball field is not far off from the green colour of that image. The resulting lateral neighbours, displayed in the left panel of Fig 16(a), do not contain the desired scene. However, there is an image of a sports field. Making that the focal image, as seen in the left part of fig 16(b), reveals it has the desired scene as a lateral neighbour. Clicking that will unearth a lot more images from baseball fields, see the right side of Fig 16(b). The network structure, a bit of lateral thinking and three mouse clicks have brought the desired result.

In the same way, and again with only three clicks, one could have started from the football image in the database overview to find "video shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at". Heesch (2005) has shown that this is no coincidence; lateral-neighbour networks computed in this way have the so-called *small world property* (Watts and Strogatz, 1998) with only 3–4 degrees of separation even for the large TRECVid (2003) database that contains keyframes from 32,000 video shots. Lateral browsing has proven eminently successful for similar queries (Heesch et al, 2003).

Geo-temporal browsing takes the idea of timelines and automatically generated maps, eg as offered in the Perseus Digital Library (Crane, 2005), a step further. It integrates the idea of browsing in time and space with a selection of events through a text search box. In this way, a large newspaper or TV news collection can be made available through browsing based on what happened where and when as opposed to by keyword only.

The interface in Fig 18 is a design study in our group that allows navigation within a large news event dataset along three dimensions: time, location and text subsets. The search term presents a text filter. The temporal distribution can be seen in lower part. The overview window establishes a frame of reference for the user's region of interest. In principle, this interface could implement new zooming techniques, eg speed-dependent automatic zooming (Cockburn and Savage, 2003), and link to a server holding a large quantity of maps such as National Geographic's MapMachine (http://plasma.nationalgeographic.com/mapmachine/ as of May 2005) with street-level maps and aerial photos.

0.6 Summary

This chapter has introduced basic concepts of multimedia resource discovery technologies for a number of different query and document types; these were the piggy-back text search, automated annotation, content-based retrieval and fingerprinting. The paradigms we have discussed include summarising complex multimedia objects such as TV news, information visualisation techniques for document clusters, visual search by example, relevance feedback and methods to create browsable structures within the collection. These exploration modes share three common features: they are

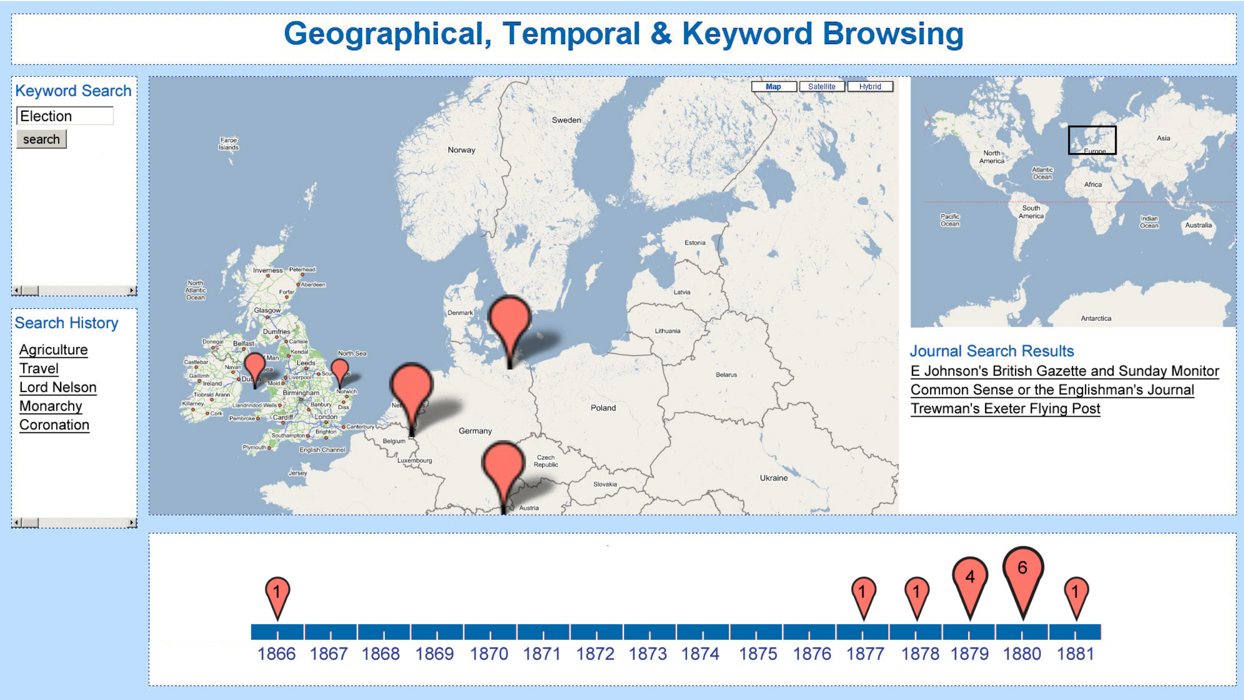


Figure 18: Geo-temporal browsing in action

automatically generated, depend on visual senses and interact with the user of the multimedia collections.

Multimedia resource discovery has its very own challenges in the semantic gap, in polysemy inherently present in under-specified query-by-example scenarios, in the question how to combine possibly conflicting evidence and the responsiveness of the multimedia searches. In the last part of the chapter we have given some examples of user-centred methods that support resource discovery in multimedia digital libraries. Each of these methods can be seen as an alternative mode to the traditional digital library management tools of meta-data and classification. The new visual modes aim at generating a multi-faceted approach to present digital content: *video summaries* as succinct versions of media that otherwise would require a high bandwidth to display and considerable time by the user to assess; *information visualisation* techniques help the user to understand a large set of documents that match a query; *visual search* and *relevance feedback* afford the user novel ways to express their information need without taking recourse to verbal descriptions that are bound to be language-specific; alternative resource discovery modes such as *lateral browsing* and *geo-temporal browsing* will allow users to explore collections using lateral associations and geographic or temporal filters rather than following strict classification schemes that seem more suitable for trained librarians than the occasional user of multimedia collections. The cost for these novel approaches will be low, as they are automated rather than human-generated. It remains to be seen how best to integrate these services into traditional digital library designs and how much added value these services will bring about (Bainbridge et al, 2005).

My 2010 book provides further reading for content-based multimedia retrieval and complements the material of this chapter: Rüger (2010), Multimedia information retrieval, Lecture notes in the series Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan and Claypool

Publishers, DOI: 10.2200/S00244ED1V01Y200912ICR010

In fact, this chapter utilised some excerpts and figures from this book.

Acknowledgements: Outlining the paradigms in this chapter and their implementations would not have been possible without the ingenuity, imagination and hard work of Paul Browne, Matthew Carey, Shyamala Doraisamy, Daniel Heesch, Peter Howarth, Suzanne Little, Hai-Ming Liu, Ainhoa Llorente, João Magalhães, Alexander May, Simon Overell, Marcus Pickering, Adam Rae, Edward Schofield, Shalini Sewraz, Dawei Song, Lawrence Wong and Alexei Yavlinsky.

Credits: The photograph in Figure 1 (Milton Keynes Peace pagoda) by Stefan Rüger, Jul 2007, was first published in (Rüger, 2010). Figure 2 is a mock-up based on the existing üBase search engine, see Fig 14, with modifications by Peter Devine and was previously published in (Rüger, 2010). Figure 3 (new search engine types) was designed by Peter Devine and published in (Rüger, 2010). Figures 5, 6 and 9 use royalty-free images from Corel Gallery 380,000, © Corel Corporation, all rights reserved. Figure 7 (Behold) by Alexei Yavlinsky are screenshots from http://photo.beholdsearch.com, 19 July 2007, now http://www.behold.cc with thumbnails of creative-commons Flickr images. The photograph in Figure 8 © by Stefan Rüger, taken May 1996 in the Århus Art Museum. Figure 8 and fig-features-distances were published in (Rüger, 2010). The screenshots in Figures 10 – 14 and 16 – 18 are reproduced courtesy of © Imperial College London. The ANSES system in Figure 10 was originally designed by Marcus Pickering and later modified by Lawrence Wong; the images and part of the text displayed in the screenshot of Figure 10 were recorded from British Broadcasting Corporation (BBC), http://www.bbc.co.uk. The Sammon map in Figure 11 and the radial visualisation in Figure 13 were designed by Matthew Carey. The Dendro map in Figure 12 was designed by Daniel Heesch. The üBase system depicted in the screenshots of Figure 14 (a), 14 (b) and 16 (a) was designed by Alexander May. The images used within the screenshot of Figure 14 and within the illustration of Figure 15 were reproduced from Corel Gallery 380,000, © Corel Corporation, all rights reserved. The images in the (partial) screenshots of Figures 16 and 16 were reproduced from TREC Video Retrieval Evaluation 2003 (TRECVid), http://www-nlpir.nist.gov/projects. The geotemporal browsing screenshot in Figure 18 was created by Simon Overell.

Bibliography

C Aggarwal and P Yu (2000). The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space. In *ACM International Conference on Knowledge Discovery and Data Mining*, pp 119–129. DOI: 10.1145/347090.347116.

L von Ahn and L Dabbish (2004). Labeling images with a computer game. In *ACM International Conference on Human Factors in Computing Systems*, pp 319–326. DOI: 10.1145/985692.985733.

M Ankerst, D Keim and H Kriegel (1996). Circle segments: a technique for visually exploring large multidimensional data sets. In *IEEE Visualization*.

J Aslam and M Montague (2001). Models for metasearch. In *ACM International Conference on Research and Development in Information Retrieval*, pp 276–284.

P Au, M Carey, S Sewraz, Y Guo and S Rüger (2000). New paradigms in information visualisation. In *ACM International Conference on Research and Development in Information Retrieval*, pp 307–309. DOI: 10.1145/345508.345610.

M Baillie and J Jose (2004). An audio-based sports video segmentation and event detection algorithm. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp 110. DOI: 10.1109/CVPR.2004.298.

D Bainbridge, P Browne, P Cairns, S Rüger and L-Q Xu (2005). Managing the growth of multimedia digital content. *ERCIM News: special theme on Multimedia Informatics 62*, 16–17.

B Bartell, G Cottrell and R Belew (1994). Automatic combination of multiple ranked retrieval systems. In *ACM International Conference on Research and Development in Information Retrieval*, pp 173–181.

J Beis and D Lowe (1997). Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp 1000–1006. DOI: 10.1109/CVPR.1997.609451.

W Birmingham, R Dannenberg and B Pardo (2006). Query by humming with the VocalSearch system. *Communications of the ACM 49*(8), 49–52. DOI: 10.1145/1145287.1145313.

D Blei and M Jordan (2003). Modeling annotated data. In *ACM International Conference on Research and Development in Information Retrieval*, pp 127–134. DOI: 10.1145/860435.860460.

K Börner (2000). Visible threads: a smart VR interface to digital libraries. In *International Symposium on Electronic Imaging 2000: Visual Data Exploration and Analysis*, pp 228–237. DOI: 10.1117/12.378899.

I Campbell (2000). Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. *Journal of Information Retrieval 2*(1), 89–114. DOI: 10.1023/A:1009902203782.

P Cano, E Batlle, T Kalker and J Haitsma (2005). A review of audio fingerprinting. *41*(3), 271–284. DOI: 10.1007/s11265-005-4151-3.

S Card (1996). Visualizing retrieved information: a survey. *IEEE Computer Graphics and Applications 16*(2), 63–67. DOI: 10.1109/38.486683.

M Carey, D Heesch and S Rüger (2003). Info navigator: a visualization interface for document searching and browsing. In *International Conference on Distributed Multimedia Systems*, pp 23–28.

A Cavallaro and T Ebrahimi (2004). Interaction between high-level and low-level image analysis for semantic video object extraction. *Journal on Applied Signal Processing 2004*(6), 786–797. DOI: 10.1155/S1110865704402157.

B Chawda, B Craft, P Cairns, S Rüger and D Heesch (2005). Do "attractive things work better"? An exploration of search tool visualisations. In *BCS Human-Computer Interaction Conference*, Volume 2, pp 46–51.

M Christel, A Hauptmann, A Warmack and S Crosby (1999). Adjustable filmstrips and skims as abstractions for a digital video library. In *IEEE Forum on Research and Technology Advances in Digital Libraries*, pp 98–104. DOI: 10.1109/ADL.1999.777702.

M Christel and A Warmack (2001). The effect of text in storyboards for video navigation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp 1409–1412. DOI: 10.1109/ICASSP.2001.941193.

A Cockburn and J Savage (2003). Comparing speed-dependent automatic zooming with traditional scroll, pan and zoom methods. In *BCS Human-Computer Interaction Conference*, pp 87–102.

I Cox, M Miller, T Minka, T Papathomas and P Yianilos (2000). The Bayesian image retrieval system, PicHunter. *IEEE Trans on Image Processing 9*(1), 20–38.

K Cox (1992). Information retrieval by browsing. In *International Conference on New Information Technology*, pp 69–80.

K Cox (1995). *Searching through browsing*. PhD thesis, University of Canberra.

G Crane (Ed) (2005). *Perseus Digital Library Project*. Tufts Uni, 30 May 2005, http://www.perseus.tufts.edu.

B Croft and T Parenty (1985). A comparison of a network structure and a database system used for document retrieval. *Information Systems 10*, 377–390. DOI: 10.1016/0306-4379(85)90042-0.

H Cunningham (2002). GATE, a general architecture for text engineering. *Computers and the Humanities 36*, 223–254. DOI: 10.1023/A:1014348124664.

R Datta, D Joshi, J Li and J Wang (2008). Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys 40*(2), 1–60. DOI: 10.1145/1348246.1348248.

D Dearholt and R Schvaneveldt (1990). Properties of Pathfinder networks. In R Schvaneveldt (Ed), *Pathfinder associative networks: studies in knowledge organization*, pp 1–30. Norwood.

S van Dongen (2000). A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands.

S Doraisamy and S Rüger (2003). Robust polyphonic music retrieval with n-grams. *Journal of Intelligent Information Systems 21* (1), 53–70.

P Duygulu, K Barnard, N de Freitas and D Forsyth (2002). Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, pp 349–354. Springer LNCS 2353. DOI: 10.1007/3-540-47979-1_7.

P Enser and C Sandom (2002). Retrieval of archival moving imagery — CBIR outside the frame? In *International Conference on Image and Video Retrieval*, pp 85–106. Springer LNCS 2383. DOI: 10.1007/3-540-45479-9_22.

P Enser and C Sandom (2003). Towards a comprehensive survey of the semantic gap in visual image retrieval. In *International Conference on Image and Video Retrieval*, pp 163–168. Springer LNCS 2728. DOI: 10.1007/3-540-45113-7_29.

S Feng, R Manmatha and V Lavrenko (2004). Multiple Bernoulli relevance models for image and video annotation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp 1002–1009. DOI: 10.1109/CVPR.2004.1315274.

R Fowler, B Wilson and W Fowler (1992). Information navigator: an information system using associative networks for display and retrieval. Technical Report NAG9-551, 92-1, Department of Computer Science, University of Texas.

J Hare and P Lewis (2004). Salient regions for query by image content. In *International Conference on Image and Video Retrieval*, pp 264–268. Springer LNCS 3115. DOI: 10.1007/b98923.

J Hare and P Lewis (2005). Saliency-based models of image content and their application to auto-annotation by semantic propagation. In *Multimedia and the Semantic Web Workshop at the European Semantic Web Conference*.

J Hare, P Lewis, P Enser and C Sandom (2006). Mind the gap: another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management and Retrieval, SPIE Vol 6073*, pp 1–12. DOI: 10.1117/12.647755.

D Heesch (2005). *The NN^k technique for image searching and browsing*. PhD thesis, Imperial College London.

D Heesch, M Pickering, S Rüger and A Yavlinsky (2003). Video retrieval using search and browsing with key frames. In *TREC Video Retrieval Evaluation*.

D Heesch and S Rüger (2003). Relevance feedback for content-based image retrieval: what can three mouse clicks achieve? In *European Conference on Information Retrieval*, pp 363–376. Springer LNCS 2633. DOI: 10.1007/3-540-36618-0_26.

D Heesch and S Rüger (2004). NN^k networks for content based image retrieval. In *European Conference on Information Retrieval*, pp 253–266. Springer LNCS 2997. DOI: 10.1007/b96895.

M Hemmje, C Kunkel and A Willet (1994). LyberWorld — a visualization user interface supporting fulltext retrieval. In *ACM International Conference on Research and Development in Information Retrieval*, pp 249–259.

P Hoffman, G Grinstein and D Pinkney (1999). Dimensional anchors: a graphic primitive for multi-dimensional multivariate information visualizations. In *New Paradigms in Information Visualisation and Manipulation in conjunction with ACM CIKM*, pp 9–16. DOI: 10.1145/331770.331775.

P Howarth and S Rüger (2005c). Trading precision for speed: localised similarity functions. In *International Conference on Image and Video Retrieval*, pp 415–424. Springer LNCS 3568. DOI: 10.1007/11526346_45.

Y Ishikawa, R Subramanya and C Faloutsos (1998). MindReader: Querying databases through multiple examples. In *International Conference on Very Large Databases*, pp 218–227.

J Jeon, V Lavrenko and R Manmatha (2003). Automatic image annotation and retrieval using cross-media relevance models. In *ACM International Conference on Research and Development in Information Retrieval*, pp 119–126. DOI: 10.1145/860435.860459.

R Korfhage (1991). To see or not to see — is that the query? In *ACM International Conference on Research and Development in Information Retrieval*, pp 134–141. DOI: 10.1145/122860.122873.

V Lavrenko, R Manmatha and J Jeon (2003). A model for learning the semantics of pictures. In *Neural Information Processing Systems*, pp 553–560.

M Lew, N Sebe, C Djeraba and R Jain (2006). Content-based multimedia information retrieval: state of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications 2*(1), 1–19. DOI: 10.1145/1126004.1126005.

C Liu, J Yuen and A Torralba (2009a). Nonparametric scene parsing: label transfer via dense scene alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp 1972–1979. DOI: 10.1109/CVPRW.2009.5206536.

J Magalhães and S Rüger (2006). Logistic regression of semantic codebooks for semantic image retrieval. In *International Conference on Image and Video Retrieval*, pp 41–50. Springer LNCS 4071. DOI: 10.1007/11788034_5.

J Magalhães and S Rüger (2007). Information-theoretic semantic multimedia indexing. In *International Conference on Image and Video Retrieval*, pp 619–626. DOI: 10.1145/1282280.1282368.

A Makadia, V Pavlovic and S Kumar (2008). A new baseline for image annotation. In *European Conference on Computer Vision*, pp 316–329. Springer LNCS 5304. DOI: 10.1007/978-3-540-88690-7_24.

D Metzler and R Manmatha (2004). An inference network approach to image retrieval. In *International Conference on Image and Video Retrieval*, pp 42–50. Springer LNCS 3115. DOI: 10.1007/b98923.

W Müller and A Henrich (2004). Faster exact histogram intersection on large data collections using inverted VA-files. In *International Conference on Image and Video Retrieval*, pp 455–463. Springer LNCS 3115. DOI: 10.1007/b98923.

S Nene and S Nayar (1997). A simple algorithm for nearest neighbor search in high dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*(9), 989–1003. DOI: 10.1109/34.615448.

M Pickering (2004). *Video Retrieval and Summarisation.* PhD thesis, Imperial College London.

M Pickering, L Wong and S Rüger (2003). ANSES: summarisation of news video. In *International Conference on Image and Video Retrieval*, pp 481–486. Springer LNCS 2728. DOI: 10.1007/3-540-45113-7_42.

K Rodden, W Basalaj, D Sinclair and K Wood (1999). Evaluating a visualization of image similarity. In *ACM International Conference on Research and Development in Information Retrieval*, pp 36–43.

Stefan Rüger (2009). Multimedia resource discovery. In Ayşe Göker and John Davies (Eds), *Information Retrieval: Searching in the 21st Century*, pp 39–62. Wiley.

Stefan Rüger (2010). *Multimedia information retrieval*. Morgan and Claypool Publishers. DOI: 10.2200/S00244ED1V01Y200912ICR010.

Y Rui, T Huang and S Mehrotra (1998). Relevance feedback techniques in interactive content-based image retrieval. pp 25–36.

J Rydberg-Cox, L Vetter, S Rüger and D Heesch (2004). Approaching the problem of multi-lingual information retrieval and visualization in Greek and Latin and Old Norse texts. In *European Conference on Digital Libraries*, pp 168–178. Springer LNCS 3232. DOI: 10.1007/b100389.

A Salway and M Graham (2003). Extracting information about emotions in films. In *ACM Conference on Multimedia*, pp 299–302. DOI: 10.1145/957013.957076.

A Salway, A Vassiliou and K Ahmad (2005). What happens in films? In *IEEE International Conference on Multimedia and Expo*, pp 4. DOI: 10.1109/ICME.2005.1521357.

J Sammon (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers 18*(5), 401–409. DOI: 10.1109/T-C.1969.222678.

S Santini and R Jain (2000). Integrated browsing and querying for image databases. *IEEE Multimedia 7*(3), 26–39. DOI: 10.1109/93.879766.

J Seo, J Haitsma, T Kalker and C Yoo (2004). A robust image fingerprinting system using the radon transform. *Signal Processing: Image Communication 19*, 325–339.

J Shaw and E Fox (1994). Combination of multiple searches. pp 243–252.

B Shneiderman, D Feldman, A Rose and X Ferré Grau (2000). Visualizing digital library search results with categorical and hierarchical axes. In *ACM Conference on Digital Libraries*, pp 57–66. DOI: 10.1145/336597.336637.

A Smeaton, C Gurrin, H Lee, K Mc Donald, N Murphy, N O'Connor, D O'Sullivan, B Smyth and D Wilson (2004). The Físchlár-news-stories system: personalised access to an archive of TV news. In *RIAO Conference on Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pp 3–17.

D Squire, W Müller, H Müller and T Pun (2000). Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters 21*(13–14), 1193–1198.

T Tolonen and M Karjalainen (2000). A computationally efficient multi-pitch analysis model. *IEEE Transactions on Speech and Audio Processing 8*(6), 708–716. DOI: 10.1109/89.876309.

A Torralba and A Oliva (2003). Statistics of natural image categories. *Network: Computation in Neural Systems 14*, 391–412. DOI: 10.1088/0954-898X/14/3/302.

TRECVid (2003). Trec video retrieval evaluation. http://www-nlpir.nist.gov/projects/tv2003/ last accessed Feb 2006.

G Tzanetakis and P Cook (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing 10*(5), 293–302. DOI: 10.1109/TSA.2002.800560.

J Voss (2007). Tagging, folksonomy & Co — renaissance of manual indexing? *Computing Research Repository abs/cs/0701072*, 1–12.

A de Vries, N Mamoulis, N Nes and M Kersten (2002). Efficient k-nn search on vertically decomposed data. In *ACM International Conference on Management of Data*, pp 322–333. DOI: 10.1145/564691.564729.

D Watts and S Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*, 440–442. DOI: 10.1038/30918.

R Weber, H-J Stock and S Blott (1998). A quantitative analysis and performance study for similarity search methods in high-dimensional space. In *International Conference on Very Large Databases*, pp 194–205.

M Wood, N Campbell and B Thomas (1998). Iterative refinement by relevance feedback in content-based digital image retrieval. In *ACM Multimedia*, pp 13–20.

A Yavlinsky, M Pickering, D Heesch and S Rüger (2004). A comparative study of evidence combination strategies. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp 1040–1043.

A Yavlinsky, E Schofield and S Rüger (2005). Automated image annotation using global features and robust nonparametric density estimation. In *International Conference on Image and Video Retrieval*, pp 507–517. Springer LNCS 3568. DOI: 10.1007/11526346.

Index